Journal of Electrocardiology 58 (2020) 143–149

Contents lists available at ScienceDirect

Journal of Electrocardiology

journal homepage: www.jecgonline.com

Performance of seven ECG interpretation programs in identifying arrhythmia and acute cardiovascular syndrome

J. De Bie, PhD [a,*], C. Martignani, PhD [b], G. Massaro, MD [b], I. Diemberger, MD PhD [b]

[a] Mortara Instrument Europe s.r.l., Bologna, Italy
[b] Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy

ARTICLE INFO

Keywords:
Acute coronary syndrome
Atrial fibrillation
Electrocardiograph interpretation programs
Myocardial infarction
Sinus rhythm

ABSTRACT

Background: No direct comparison of current electrocardiogram (ECG) interpretation programs exists.
Objective: Assess the accuracy of ECG interpretation programs in detecting abnormal rhythms and flagging for priority review records with alterations secondary to acute coronary syndrome (ACS).
Methods: More than 2,000 digital ECGs from hospitals and databases in Europe, USA, and Australia, were obtained from consecutive adult and pediatric patients and converted to 10 s analog samples that were replayed on seven electrocardiographs and classified by the manufacturers' interpretation programs. We assessed ability to distinguish sinus rhythm from non-sinus rhythm, identify atrial fibrillation/flutter and other abnormal rhythms, and accuracy in flagging results for priority review. If all seven programs' interpretation statements did not agree, cases were reviewed by experienced cardiologists.
Results: All programs could distinguish well between sinus and non-sinus rhythms and could identify atrial fibrillation/flutter or other abnormal rhythms. However, false-positive rates varied from 2.1% to 5.5% for non-sinus rhythm, from 0.7% to 4.4% for atrial fibrillation/flutter, and from 1.5% to 3.0% for other abnormal rhythms. False-negative rates varied from 12.0% to 7.5%, 9.9% to 2.7%, and 55.9% to 30.5%, respectively. Flagging of ACS varied by a factor of 2.5 between programs. Physicians flagged more ECGs for prompt review, but also showed variance of around a factor of 2. False-negative values differed between programs by a factor of 2 but was high for all (>50%). Agreement between programs and majority reviewer decisions was 46–62%.
Conclusions: Automatic interpretations of rhythms and ACS differ between programs. Healthcare institutions should not rely on ECG software "critical result" flags alone to decide the ACS workflow.

Introduction

Computer programs designed to generate and interpret electrocardiograms (ECGs) have been available for >50 years. Their use has spread since the 1980s, when real-time analysis and direct print of the results on the ECG were introduced. Improvements in automatic ECGs analysis have shifted its role from saving the time of cardiologists to supporting diagnosis when access to a specialist is not possible. However, owing to program shortcomings, artifacts, and recording errors, results still warrant being over-read by experienced clinicians, in particular in the clinical context [1]. Nevertheless, although not intended by equipment manufacturers, ECGs have become increasingly interpreted by less experienced physicians who rely more heavily on the computer interpretations and measurements. Additionally, efficiency requirements, automation, and electronic workflows have separated ECG acquisition from centralized assessment by cardiologists, often creating delays before a confirmed report becomes available for clinical decision making.

The quality of outputs by ECG interpretation programs has been consistently questioned. Some studies have assessed clinical accuracy [2], but have assessed only single programs and used ECGs available only to the researchers [3], as is done by manufacturers to validate their programs. Study results, therefore, are not comparable.

A few direct comparisons of multiple programs using the same set of ECGs were performed in the 1970s to 1980s. Meyer et al. [4,5] compared six programs on around 250 ECGs in 1974, MacFarlane et al. [6] compared two programs on 300 ECGs in 1981, and Bjerle and Niklasson [7] used 200 ECGs to assess three unnamed programs in 1988. These numbers are small compared with the sets of ECGs normally used to develop a reference set based on cardiologists' opinions. Additionally, results from these studies are out of date because they include programs reading ECGs based on three orthogonal (Frank) leads, whereas 12-lead ECGs are now used exclusively; also, the programs tested are no longer available.

In 1991, Willems et al. [8] compared interpretations of 1220 Frank and 12-lead ECGs by eight cardiologists with statements from nine

* Corresponding author at: Mortara Instrument Europe s.r.l., Via Cimarosa 103, 40033 Casalecchio di Reno, Bologna, Italy.
E-mail address: Johannes.DeBie@hillrom.com (J. De Bie).

https://doi.org/10.1016/j.jelectrocard.2019.11.043
0022-0736/

programs. The programs clearly differed, but the outcomes had limited value for everyday clinical use as only ECGs from patients with ventricular hypertrophy and/or old myocardial infarction were used. We found no other studies published in the past 25 years that directly compared ECG reading programs currently on the market. The clinical management of atrial fibrillation, in particular, may be effectively improved with the use of reliable home- or office-based ECG analysis [9].

ECG program users have few real-world data to aid in the choice of program other than personal experience. In this study, we compared interpretations of seven mainstream programs in a large set of ECGs that reflect everyday clinical use. We report performance regarding interpretation statement accuracy in identification of arrhythmia and the ability of programs to flag acute coronary syndrome (ACS) that requires prompt review by a clinician. These are the categories where automatic interpretation most impacts clinical decision making in practice.

Methods

Selection of ECGs

For the ACS and arrhythmia analyses, we used a set of anonymized ECGs acquired consecutively in adults and children in eight hospitals and acute-care centers in the USA, Italy, and Australia. We also added to the arrhythmia a set of ECGS from a European ambulance service and university hospital that had critical value statements flagging acute myocardial infarction by one of the programs. ECGs from patients with pacemakers were excluded (Supplementary Methods).

Re-recording of ECGs

ECGs were converted into analog format for replay into physical electrocardiographs, and 10 s 1000 samples/s looped records were created to enable feeding into electrocardiographs. Precautions were taken to avoid discontinuity, create an RR interval that was the average of the whole record and ensure that all records were exactly 10 s in length (Supplementary Methods and Fig. S1). We excluded records that needed to be stretched or compressed by >10% or had a large amplitude discontinuity.

Analog records were replayed in continuous 10 s loops with a Whaleteq MECG 2.0 Multichannel ECG Test System (WHALETEQ Co Ltd., Taipei City, Taiwan) connected to a laptop PC. Of the seven electrocardiographs tested, which were labelled A–G, up to four were connected in parallel during recording (Supplementary Methods). The details of the electrocardiographs are shown in Table 1. Printed ECGs were visually checked periodically to ensure that they matched the source record. The electrocardiographs were configured to provide full ECG interpretations in English with filters set to a minimum or off. Each ECG was recorded three times on every electrocardiograph.

Because ECG capture could not be guaranteed to start in the same place of the continuous loop every time, and because of small reproduction discrepancies, amplifier noise, and sampling effects, we expected slight differences within sets of three ECGs recordings. Therefore, we selected the most representative or the least pathological interpretation of the three recordings for the analysis (Supplementary Methods).

Manufacturer interpretation statements

Each manufacturer uses different wording for interpretation statements, including those for probability and severity of conditions. Thus, to render individual statements independent from specific wording, we created classes of statement types, causes, and locations. For the arrhythmia analysis we grouped the classifications into three categories: “sinus rhythm”, “atrial fibrillation/flutter”, and “other arrhythmias”. Heart rate, which programs reliably detect and calculate [10], did not affect our rhythm classifications (tachycardia and bradycardia were grouped together with normal rate). For the ACS analysis, we included interpretation statements that resulted in a “critical value” or “critical test result” or, for manufacturers who did not support such statements, those indicative of possible ACS (e.g., “acute MI”; Table 1). Text parsing rules were created for the infarction interpretations, resulting a true/false decision for ACS.

Table 1
Electrocardiographs used to acquire ECGs and critical value interpretation statements for ACS

ACS statements	Critical value statement
GE Healthcare MAC2000, 12SL program version 22[a]	
[location] infarct, possibly acute[b]	** **Acute MI/stemi** **
[location] injury pattern[b]	** **Acute MI/stemi** **
ST elevation, consider [location] injury or acute infarct[b]	** **Acute MI/stemi** **
Glasgow Burdick 8500 program version 26.5[c]	
Acute [location] infarct[d]	***Consider acute stemi***
[location] infarction – possibly acute[d]	***Consider acute stemi***
Strongly suggests myocardial injury/ischemia	***Acute MI/ischemia***
Strongly suggests myocardial infarction	***Acute MI/ischemia***
Consider acute infarction	***Acute MI/ischemia***
WelchAllyn CP150, MEANS program, Revision 2016–7	
Consider infarct of recent occurrence	Not available
Consider infarct of acute occurrence	Not available
Consider acute ischemia	Not available
Consider pericarditis	Not available
Midmark IQ-manager resting ECG, Interpretation program version 8.6.1	
[size] [location] infarct, [probability] recent[e,f]	Not available
[location] ST-elevation, consider acute process[b]	Not available
[location] Injury[b]	Not available
Mortara Instrument ELI 380, Veritas Interpretation program version 7.3.0	
[location] myocardial infarction, probably recent[b]	***Acute MI***
[location] myocardial infarction, possibly acute[b]	***Acute MI***
[severity] ST elevation, consider [location] injury[g]	***Acute MI***
[probability] acute pericarditis - exclude acute mi[h]	***Acute MI***
Marked ST depression, consider subendocardial injury	***Acute MI***
Philips TC 20, DXL Interpretation program version PH100B	
[probability] [location] infarct – acute[b]	>>>Acute MI<<<
[probability] [location] infarct – possibly acute[b]	>>>Acute MI<<<
Repol ABNRM - severe global ischemia (LM/MVD)[i]	>>>Acute ischemia<<<
Schiller MS2015, Interpretation program R16.01	
Consistent with [location] infarct, possibly recent[b]	Not available
Consider recent [location] myocardial or pericardial damage[b]	Not available
Consider acute [location] infarct[b]	Not available

[a] An optional analysis program ACS tool is also provided, which assumes that symptoms commensurate with acute coronary syndrome are also present and generates additional critical value statements related to ACS, but this program was not used.
[b] [location] can be any combination of ECG locations.
[c] A “pericarditis” statement is available but does not lead to a critical value.
[d] [location] can be any combination of ECG locations or “extensive”.
[e] [size] can be empty or “extensive”.
[f] [probability] can be “possible” or “probable”.
[g] [severity] can be “marked” or empty.
[h] [probability] can be “possible” or empty.
[i] [severity] can be “probable”, “extensive” or empty.

ECG interpretation

In the rhythm analysis, if all seven programs agreed, interpretations did not need expert review. If any program disagreed, the ECG was presented without interpretation statements to an independent experienced cardiologist for review and the interpretation was confirmed by JdB (Supplementary Methods). If interpretations did not match, the case was discussed until consensus was reached. In order to check our hypothesis that no review was needed when all programs agreed, we also reviewed a subset of those ECGs (all 141 abnormal rhythm ECGs and the same number of randomly selected sinus rhythm ECGs).

For the ACS analysis, if any program disagreed with the interpretations of others, the ECG was reviewed separately by three experienced cardiologists and the majority interpretation was used for the analysis. They were asked to judge whether a possible ongoing acute cardiovascular event was indicated and whether in practice a technician should consult a clinician immediately or process the ECG normally, knowing that it could take several hours before a physician would see it. The reviewers were asked to ignore concurrent reasons for priority processing, such as severe arrhythmias. They also reviewed the ten positive ACS cases where all programs agreed and 53 cases where any program assigned a "subendocardial ischemia" or "pericarditis" statement without leading to a "critical result".

Statistical analysis

Program performance was measured by the false-positive rate (i.e., would incorrectly trigger prompt physician consultation), calculated as the number of false-positive results divided by the total number of negative cases, and the false-negative rate, calculated as the number of missed abnormal cases divided by all abnormal cases. We used the Wilson score to calculate 95% confidence intervals (CIs). To assess the probability that the performance did not differ between programs, we used the McNemar test to compare individual scoring on the same dataset. We took $p < 0.1$ to be significant. All significant p-values are reported, whereas non-significant p-values are not reported.

Cohen's kappa coefficient (κ) was used to measure inter-rater agreement for qualitative (categorical) items as a metric of overall agreement between programs and expert reviewers. We calculated 95% CIs for κ values based on the standard error, as described by Fleiss et al. [11]

We did two exploratory analyses. First, we investigated how programs performed in identification of the arrhythmia categories when heart rate was <100 versus ≥100 beats per min. Second, we assessed whether disagreement between programs about the presence of ACS in ECGs was increased with increasing QRS duration, only for ECGs from adults to avoid bias towards lower QRS durations.

Results

We replayed 2610 ECGs for the rhythm analysis and 2382 for the ACS analysis, of which 2155 and 1986 remained after exclusions for stretching or compression and spline slope. Twenty-six records were further excluded due to technical quality and six records were discarded because the operator had played back the wrong record for some programs. Therefore, 2123 ECGs were used in the rhythm analysis. As the ECGs from the ambulance service and university hospital already had critical test results flagged by a single program, we excluded those from the ACS analysis to avoid selection bias, meaning that 1954 were used in this analysis. We were able to parse and extract identifiers, global ECG measurements and interpretation statements for all records.

Rhythm statements

Of 2123 ECGs, all seven programs agreed that 1645 (77.9%) showed sinus rhythm, 139 cases (6.5%) showed atrial fibrillation/flutter, and three (0.1%) showed other abnormal rhythms. In 346 cases (16.3%), at least one program did not agree with the others (including 10 where the difference was only between atrial fibrillation and flutter) and these cases were reviewed by cardiologists. Reviewers interpreted 237 as sinus rhythm, 53 as atrial fibrillation/flutter, and 56 as other rhythms. Thus, in the overall set of 2123 ECGs, 1881 (88.6%) showed sinus rhythm, 183 (8.6%) atrial fibrillation/flutter, and 59 (2.8%) other abnormal rhythms.

For distinguishing between sinus and non-sinus rhythm, false-positive rates ranged from 2.1% to 5.5% (program G, $p < 0.002$ for difference from all other programs; Fig. 1A). The worst false-negative rate was 12.0% for program G, which differed significantly from programs C and E (both 92.5% and $p = 0.05$) and program D (7.9%, $p < 0.1$; Fig. 1B). Overall agreement with reviewers ranged from $\kappa = 73\%$ to $\kappa = 86\%$ (Fig. 1C). For the 282 reviewed cases in which the seven programs agreed (141 with non-sinus rhythm, 141 with sinus rhythm), we found 100% agreement between the programs and the expert reviewers, confirming our assumption that review was not needed where all programs agreed.

Most programs identified atrial fibrillation/flutter with false-positive rates below 2% and false-negative rates <10% (Fig. 1D and E). The lowest false-positive rate (program A, 0.7%) differed significantly from for all other programs except program C ($p = 0.02$ versus program D and $p < 0.0006$ versus each of the other programs), as did the highest false-positive rate (program G, 4.4%; all comparisons $p < 0.00001$). The lowest false-negative rate was 2.7% for program E, which differed from four other programs (programs D and G $p < 0.003$, programs A and F $p < 0.09$; Fig. 1E). Overall agreement between programs and cardiologist reviewers was best for program A and C, closely followed by the other programs except program G, which lagged behind (Fig. 1F).

For other abnormal rhythms, false-positive rates ranged from 1.5% to 3.0% (Fig. 1G). The lowest rate (program E) differed significantly from three other programs ($p < 0.004$ versus programs C and D, $p < 0.04$ versus program F) and the highest rate (program D) differed significantly from four programs ($p < 0.008$ versus program A and E and $p < 0.03$ versus programs B and G). The false-negative rate was high, ranging from 55.9% to 30.5% (Fig. 1H), with the lowest value for program D, differing significantly from almost all other programs ($p < 0.002$ versus programs B and G, $p < 0.02$ versus programs C and F, and $p = 0.08$ versus program E). Overall agreement was poor for all programs (Fig. 1I).

ECGs with heart rate ≥100 bpm were, as expected, more likely to be atrial fibrillation/flutter than those with lower rates (Table 2). Also as expected, programs differed more in interpretation, with all seven agreeing in only 72% of cases of atrial fibrillation/flutter in the setting of high heart rate. False-positive and false-negative rates increased with high heart rate.

Out of curiosity, we counted the number of cases with abnormal rhythm where at least one of the programs gave the conclusion "normal ECG" (including "otherwise normal", "probably normal", and "normal except for rate"). Of the 242 ECGs showing abnormal rhythms, for 11 at least one program gave reported "normal ECG", and in four cases two or more programs did so. All these cases were ectopic atrial or junctional rhythms, with small or difficult to detect P waves.

ACS

All seven programs agreed that ACS was not present in 1747 (89.4%) of 1954 cases and that ACS was present in 10 cases (0.5%). In the remaining 197 cases (10.1%), at least one program differed from the others.

The highest frequency of flagging ACS was more than double the lower frequency (Fig. 2). Reviewers were more inclined to flag an ECG for prompt review than any of the programs, but also individually differed from each other up to a factor of two (Fig. 2). The program that flagged ACS most frequently was only third highest for agreement between programs and the majority reviewer interpretation (Fig. 3). The greatest agreement was 62% (program E) and was significantly greater than four other programs ($p < 0.05$ versus programs B, D, F and G) and least was 46% (program G), which was significantly lower than three other programs ($p < 0.01$ versus programs A, C and E). The ACS flagging frequency was lowest for program A (Fig. 2), but the same program was second highest in terms of agreement with majority cardiologist review (56%; Fig. 3).

Using the majority reviewer judgement as the reference standard, the highest ACS false-positive rate was 1.6% for programs B and F ($p < 0.04$ versus programs A, C, E and GG; Fig. 4A). One program (program A) had almost no false positives (0.1%; $p < 0.0005$ for all comparisons with other programs). The lowest false-

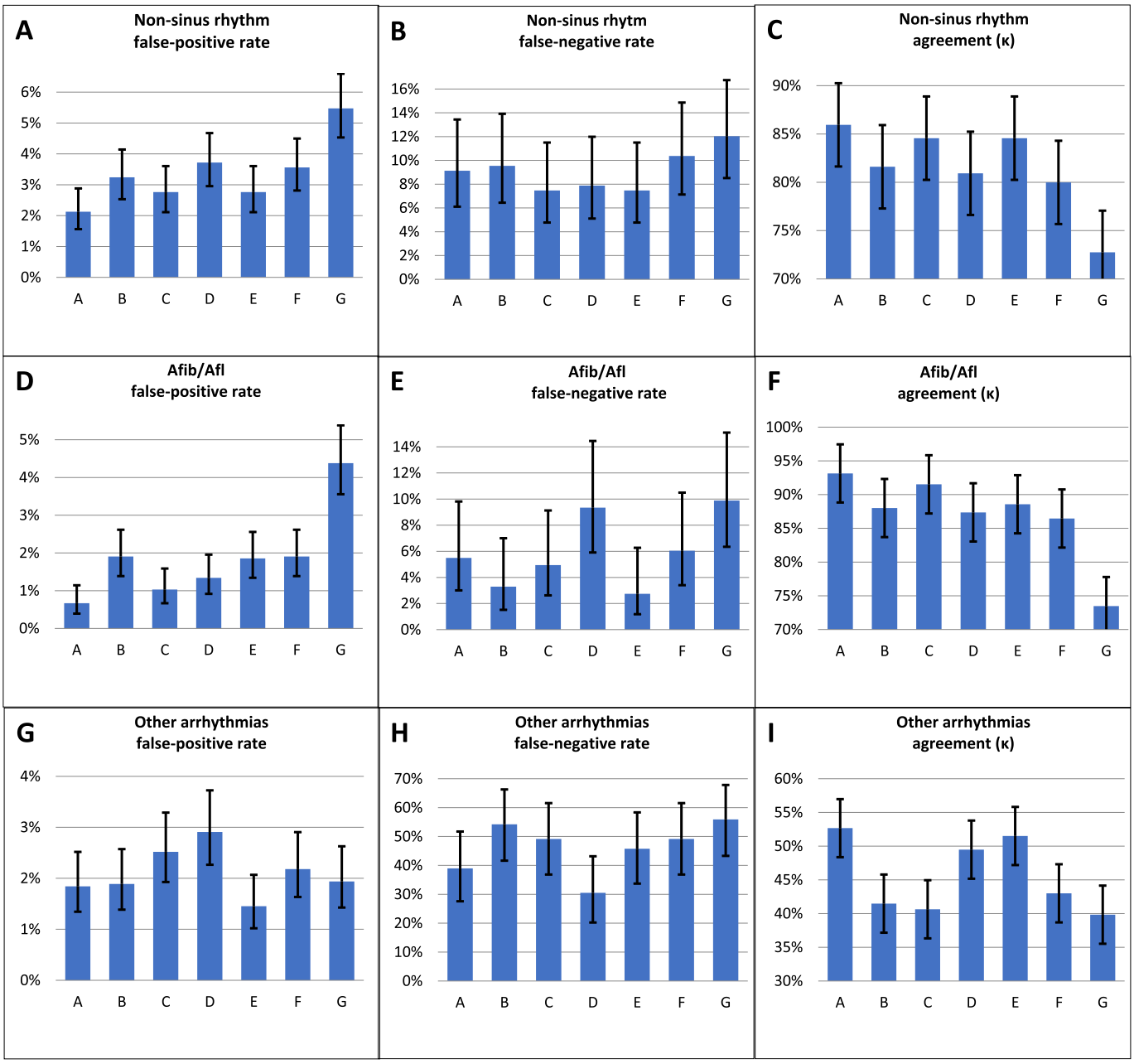

Fig. 1. Program performance in automatic rhythm interpretation. A–C: Sinus versus non-sinus rhythm. D–F: Afib/afl versus no Afib/fl. G–I: Other abnormal rhythms. False-positive rate is the number of false-positive results divided by the total number of normal cases. False-negative rate is the number of captured abnormal cases divided by all abnormal cases. 95% confidence intervals are calculated with the Wilson score method. Inter-rater agreement was measured with Cohen's kappa coefficient (κ). Abbreviations: Afib/afl = atrial fibrillation/flutter.

negative rate (programs B, E and F) was less than half the highest value (program G, $p < 0.001$) and was significantly lower than the three programs with intermediate values (programs A, C and D, all $p < 0.02$). All false-negative values were above 50% (Fig. 4B). Accordingly, overall agreement with reviewers was poor and variable (Fig. 4C).

All three reviewers agreed with the program interpretations for the 10 cases showing ACS. Fifty-three ECGs for which some programs issued a "subendocardial injury" or "pericarditis" statement without leading to a "critical result" were reviewed by the three cardiologists. The majority interpretations indicated that 18 (34.0%) of these cases warranted prompt review.

Table 2
Distribution of ECGs and program performance by heart rate and rhythm.

Heart rate	Total number	Sinus rhythm		Atrial fibrillation/flutter			
		Present	All programs agree	Present	All programs agree	Average false-positive rate	Average false-negative rate
<100 beats per min	1783	91.3%	89.1%	6.8%	78.5%	1.6%	5.3%
≥100 beats per min	340	74.7%	76.4%	17.9%	72.1%	3.5%	7.3%

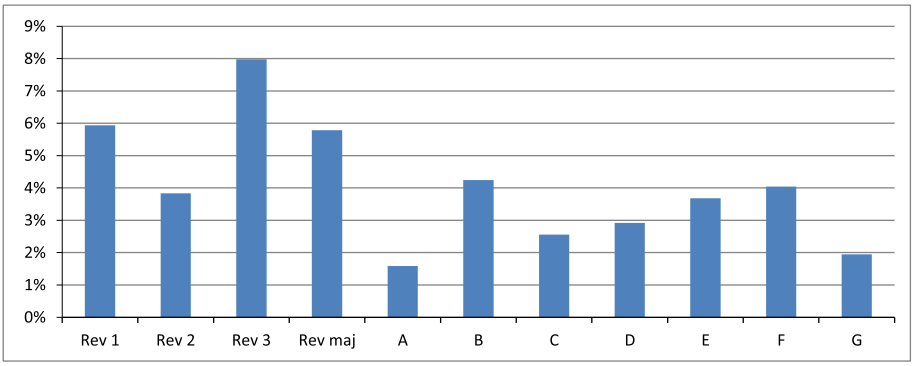

Fig. 2. Frequency of reviewer decisions or statements for programs A–G flagging possible acute coronary syndrome. Rev. = reviewer. Rev. maj = majority reviewer decision.

When ordered by QRS duration, curiously, programs increasingly indicated ACS as QRS duration increased (Table 3) while, cardiologists tended to classify more ECGs as requiring prompt review in the intermediate QRS duration categories.

Again, out of curiosity, we counted the number of cases where the cardiologist indicated prompt review of the ECG and at least one program gave indicated "normal ECG" (including "otherwise normal", "probably normal", and "normal except for rate"). Of the 122 cases, for 10 at least one program gave the conclusion "normal ECG", and in four cases two programs did so. One program (D) was responsible for nine of 10 cases. Six of the 10 cases showed slight inferior ST-elevation below AHA/ACC/ESC criteria for STEMI, but enough to make the cardiologists suspicious and for some of the programs to flag the ECGs.

Discussion

Our method allowed objective comparison of ECG program rhythm interpretation and performance in terms of indicating ECGs that cardiologists believe require prompt consultation because of suspected ACS.

Rhythm interpretation performance

Correct determination of the principal rhythm on resting ECG is diagnostic and, therefore, is an important feature of an interpretation program. Although performance for rhythm interpretation is relatively easy to measure, our database was too small to assess performance reliably in a wide range of separate arrhythmia categories. Thus, we assessed performance only in terms of identifying atrial fibrillation/flutter, sinus versus non-sinus rhythms, and all other abnormal rhythms grouped together. All programs except one distinguished sinus from non-sinus rhythm with low false-negative rates and a false-positive rate of 3.5% or less (κ = 80–86%). The worst program had both the highest false-negative rate (12.0%) and the highest false-positive rate (5.5%, κ = 73%) and was also worst in detecting atrial fibrillation/flutter (κ = 73% compared to 86–93% for the other programs). Manufacturers have clearly made different choices when balancing false-positive against false-negative rates. It could be argued that false-negative rates have a higher clinical impact than false-positive rates, but in practice, due to low prevalence of the disease and the lack of expert reviewers in some situations, false-positive values can lead to mistreatment. All programs were clearly less sensitive for other abnormal rhythms than for atrial fibrillation/flutter, ranging from 46% to 69%, and only three programs reached $\kappa \geq 50\%$. As expected, high heart rate led to decreasing agreement amongst programs for a diagnosis of atrial fibrillation/flutter.

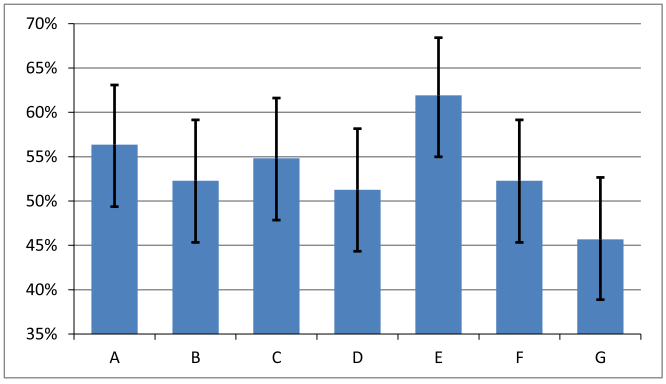

Fig. 3. Proportions of ECGs for which the majority reviewer vote and program agreed on prompt evaluation for ACS. Vertical axis is the fraction of the reviewed ECGs in which the reviewers and the program agreed. 95% confidence intervals are calculated with the Wilson score method. ACS = acute coronary syndrome.

Our methodology to determine rhythm analysis performance was similar to that proposed by the Common Standards in Electrocardiography (CSE) working group in the late 1980s and included in the international performance standard for electrocardiographs IEC 60601-2-51 in 2003 [12]. When the standard was replaced in 2011 with IEC 60601-2-25 [13], specific methods for interpretation performance testing were removed, no specific database was suggested, and no guidance was given on how to create a database and avoid bias. Minimum numbers of ECGs needed to assess sinus rhythm and atrial fibrillation were defined (1500 and 100, respectively), but for other categories the numbers were vague. Our results indicate that, although the recommended numbers might be sufficient to demonstrate minimum performance, they are too low to compare current state-of-the art programs. Our database included 192 cases of atrial fibrillation/flutter, but all programs agreed in 139 cases, leaving only 53 cases for a differentiating comparison. Similarly, in 1645 of 1882 cases of sinus rhythm, all seven programs agreed, leaving only 237 cases for comparison. We hope that this work will stimulate interest in creating a large open-access database of annotated 12-lead resting ECGs and developing a suitable platform to enable objective comparative testing of ECG interpretation program performance.

The need for an atrial fibrillation screening program has been expressed due to high associated morbidity and mortality and increased incidence and prevalence in the ageing population [14]. Several approaches have been proposed, but the most effective needs to be identified [9,15,16]. Currently, 12-lead ECG still represents the gold standard for diagnosis of atrial fibrillation [9]. Given the low false-negative and low false-positive rates we found for most programs, automatic identification of atrial fibrillation on 12-lead ECGs with review by a cardiologist of only the positive results could represent a useful approach, especially in subgroups with increased prevalence of this kind of arrhythmia (e.g., elderly patients).

ACS interpretation performance

ECG ACS statements are not diagnostic by themselves, and must be used in combination with symptoms, other clinical observations, and

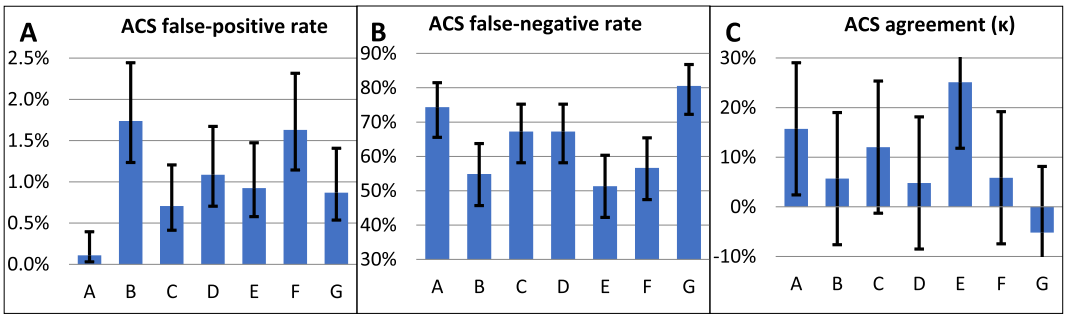

Fig. 4. Performance of programs in identifying prompt evaluation needed for ACS compared with the majority viewer vote. Data are calculated for all ECGs, either reviewed or agreed on by all programs. 95% confidence intervals are calculated with the Wilson score method. Inter-rater agreement was measured with Cohen's kappa coefficient (κ). Abbreviations: ACS = acute coronary syndrome.

tests to inform the diagnosis. As we did not include other tests, we could not compare the capability of programs to contribute effectively to the diagnosis of ACS. Instead, we assessed whether interpretation programs could avoid delays in the clinical workflow by flagging priority cases for physician review. We based the need for review on programs' "critical result" statements or an interpretation of acute MI. The ECG interpretation programs fell into three groups for level of false-negative rates compared with the majority reviewer judgements: three programs had low, two high, and two intermediate false-negative values. A low false-negative value is preferred in clinical practice, preferably without a high false-positive rate. Of the three programs with a lower false-negative rates, two also had false-positive rates >1.5% while the other (E) achieved a false-positive rate <1.0%. The program with the highest false-negative value (program A, 75%) does stand out for having almost no false-positive results.

We found notable disparity between physicians. Our reviewers were all experienced cardiologists. The most and least sensitive worked in the same department, whereas the third was from another continent. This difference in behavior, therefore, seems likely to be due to personal cautiousness or factors other than skill and experience. McCabe et al. [17] also found poor physician agreement in the absence of clinical information. Although this finding was not the principal objective of our research, it is worth further investigation.

Our results also highlight the disparity between clinicians and programs in when they recommend ECGs for priority processing. Even the most sensitive programs did not achieve false-negative values <50% compared with the majority reviews. Review of the 53 additional cases with interpretations of ST-depression or ST-elevation that did not trigger a critical program interpretation indicate that the average false-negative rate would have been even higher had we reviewed all ECGs instead of only those for which programs differed. Institutions should not solely rely on "critical result" or equivalent statements from automatic interpretation programs to prioritize cases for review. Given the substantial impact of delaying diagnosis and therapy of ACS [18,19], it is important that the "critical result" thresholds of ECG interpretation programs are more aligned with expert cardiologist opinions. Manufacturers should tune the criteria used for critical statements to improve correlation with expert opinion of when ECGs should be promptly reviewed.

Table 3
Comparison of reviewers and programs for ACS by QRS duration.

QRS duration	Total number	ACS by majority reviewer vote	ACS conclusion for at least one program
<95 ms	771	5.1%	8.6%
95–105 ms	356	10.1%	12.6%
105–120 ms	187	15.5%	21.9%
≥120 ms	129	6.2%	27.1%

Abbreviation: ACS = acute coronary syndrome.

Limitations of the study

Manufacturers do not make their programs available for comparative performance studies using stored datasets. Therefore, we needed to convert digital ECG samples to analog voltages and replay these into electrocardiographs. Small differences in signal processing, filtering and characteristics of the electronic amplification and sampling circuitry might have influenced our results. However, the effects are likely to have been small, since all electrocardiograph manufacturers adhere to internationally accepted minimum performance standards [13], and our reproductions are likely to have been close to the original ECGs. In addition, we disabled all filtering that could have influenced ECG reproduction.

Rhythm performance might have been influenced by the phase of the 10 s loop in which each ECG was captured. However, selecting the most representative interpretation of three recordings minimized this effect.

In our efforts to avoid any selection bias, we took only consecutive ECGs from several hospital databases. Although this approach results in a low number of ECGs that would be difficult to interpret (about 15% for arrhythmia and 10% for ACS), the set was representative of day-to-day workload in hospitals. To increase statistical power for comparison, a set with more abnormal or difficult cases would be needed but it would not reflect practice.

Our cardiologist panel did not review the ECGs for which none of the programs gave a critical indication for ACS. Given the low agreement on this subject between cardiologists and programs (and indeed between cardiologists themselves), there is no doubt that, had they reviewed all ECGs, the number of cases where the cardiologists flagged the ECG for prompt review would increase, resulting in even higher false-negative rates for the programs for this indication. Nevertheless, our conclusions would not have been altered.

Conclusions

This is the first time in decades that multiple current and widely used ECG interpretation programs have been directly compared using a large representative set of real-world ECGs. Notably, we developed a new methodology that can be applied without the cooperation of program manufacturers, although it is rather labor intensive. We found considerable differences between programs in interpretation, both in the ability to determine abnormal rhythms and in flagging possible ACS. For atrial fibrillation/flutter, most programs combined low false-negative and false-positive rates. For ACS, cardiologists are more inclined than programs to flag ECGs for prompt review, but false-negative rates were variable for programs and human readers. Critical test results or acute infarction statements should not be used as the sole criterion for priority processing.

Acknowledgements

The authors wish to thank Dr. Howard Cohen MD, Dr. Cristian Martignani MD, and Dr. Giulia Massaro MD for their kind review of the ECGs in this project.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary methods

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jelectrocard.2019.11.043.

References

[1] Smulyan H. The computerized ECG: friend and foe. Am J Med 2019;132:153–60.
[2] Clark EN, Sejersten M, Clemmensen P, Macfarlane PW. Automated electrocardiogram interpretation programs versus cardiologists' triage decision making based on teletransmitted data in patients with suspected acute coronary syndrome. Am J Cardiol 2010;106:1696–702.
[3] Massel D, Dawdy JA, Melendez LJ. Strict reliance on a computer algorithm or measurable ST segment criteria may lead to errors in thrombolytic therapy eligibility. Am Heart J 2000;140:221–6.
[4] Meyer J, Heinrich KW, Merx W, Effert S. Computeranalyse des elektrokardiogramms mit verschiedenen programmen I: Formanalyse. Dtsch Med Wschr 1974;99:1213–24.
[5] Meyer J, Heinrich KW, Merx W, Effert S. Computeranalyse des elektrokardiogramms mit verschiedenen programmen II: Rhythmusanalyse. Dtsch Med Wschr 1974;99:1294–300.
[6] MacFarlane PW, Melville DI, Horton MR, Bailey MD. Comparative evaluation of the IBM (12-lead) and royal infirmary (orthogonal three-lead) ECG computer programs. Circulation 1981;63:354–9.
[7] Bjerle P, Niklasson U. Comparison between three different stand-alone ECG interpretation systems. J Electrocardiol 1988;21:S163–8 suppl.
[8] Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. N Engl J Med 1991;325:1767–73.
[9] Kirchhof P, Benussi S, Kotecha D, et al. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur Heart J 2016;37:2893–962.
[10] Kligfield P, Badilini F, Denjoy I, et al. Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs. Am Heart J 2018;200:1–10.
[11] Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969;72:323–7.
[12] International Electrotechnical Committee. IEC 60601-2-51:2003: Medical electrical equipment - part 2–51: Particular requirements for safety, including essential performance, of recording and analysing single channel and multichannel electrocardiographs. Geneva: International Electrotechnical Commission; 2003.
[13] International Electrotechnical Committee. IEC 60601-2-25:2011: Medical electrical equipment - part 2–25: Particular requirements for the basic safety and essential performance of electrocardiographs. Geneva: International Electrotechnical Commission; 2011.
[14] Freedman B, Camm J, Calkins H, et al. Screening for atrial fibrillation: a report of the AF-SCREEN international collaboration. Circulation 2017;135:1851–67.
[15] Fitzmaurice DA, McCahon D, Baker J, et al. Is screening for AF worthwhile? Stroke risk in a screened population from the SAFE study. Fam Pract 2014;31:298–302.
[16] Jacobs MS, Kaasenbrood F, Postma MJ, van Hulst M, Tieleman RG. Cost-effectiveness of screening for atrial fibrillation in primary care with a handheld, single-lead electrocardiogram device in the Netherlands. Europace 2018;20:12–8.
[17] McCabe JM, Armstrong EJ, Ku I, et al. Physician accuracy in interpreting potential ST-segment elevation myocardial infarction electrocardiograms. J Am Heart Assoc 2013;2:e000268.
[18] Roffi M, Patrono C, Collet JP, et al. 2015 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: task force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2016;37:267–315.
[19] Ibanez B, James S, Agewall S, et al. 2017 ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: the task force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 2017;39:119–77.